AN EMPIRICAL CONSIDERATION OF THE USE OF R IN ACTIVELY CONSTRUCTING SAMPLING DISTRIBUTIONS

By

BRANDON K. VAUGHN

Assistant Professor, The University of Texas at Austin.

ABSTRACT

In this paper, an interactive teaching approach to introduce the concept of sampling distributions using the statistical software program, R, is shown. One advantage of this approach is that the program R is freely available via the internet. Instructors can easily demonstrate concepts in class, outfit entire computer labs, and/or assign the activity for students at home. Another benefit is that the students build their understanding of the concept in stages and are not overwhelmed by massive amounts of information all at once. By modifying the demonstrated approach, they can gradually build a conceptual foundation for sampling distributions interactively. The approach is detailed, with an empirical study considering its use in the classroom.

Keywords : Sampling Distributions, Statistical Software Program.

INTRODUCTION

For some, the use of computer technology in instruction might be without question. Many instructors now use software to help analyze real data, while others use software to demonstrate concepts covered in class. Garfield (1995) provides a wealth of resources for how students learn statistics. Computers are one proven method that has been shown to improve students' learning of data analysis. The visual aspect that computers provide students easily allow for active-learning opportunities and constructivist applications. Yet, most instructors and schools struggle with how to incorporate this. The increasing cost of popular statistical software makes it difficult for most educators to obtain and use. While some innovative statistical tutorial software has been developed in the past (e.g., ActivStats (Velleman 1996), Constats (Cohen and Chechile 1997), and so on), some of these programs are no longer updated and able to run on modern computers. And similar to statistical software, this solution results in an extra cost to students or schools to implement. Free online JAVA applets have been used as well, yet these applets are not customizable to the average user. Some resort to graphing calculators as a cheaper alternative, even though in practice one might hardly use such a device for actual statistical analysis. Graphing calculators are often limited in terms of their visual capabilities and harder to program in such a way that would enable customized use per instructor. In fact, this is a typical problem with most of the above approaches the ability to customize the activity around the uniqueness of the class and instructor.

One solution is the statistical language R, which is freely available on the internet (http://www.r-project.org). By using R, instructors can outfit an entire computer lab or easily have students download it at home. The program allows for great flexibility and customization for instruction unlike some statistical programs or computer programming (e.g., JAVA). For example, the typical statistical instructor may not have knowledge of JAVA programming, but could easily construct an R approach to most JAVA applets that are available. The statistical and graphical capabilities of R either rival or surpass those of other software packages. Furthermore, the program is available for both Windows and Mac operating systems, unlike most commercial statistical or tutorial software. However, today's society has grown accustomed to programs with a “point & click” GUI. The program R can possibly scare new users due to its command interface. Furthermore, the help files for R are at times cryptic and hard to understand, although some good textbooks on R have been published in the last few years (e.g., Crawley, 2005; Verzani, 2005; and so on). This article provides an illustration of how to use R to teach sampling distributions and concepts related to the Central Limit Theorem that is both visual and active in learning.

The purpose of this research is to compare different instructional models which involve various computer based helps and active/cooperative learning in statistics classes. Comparisons were made with previous semesters in which these approaches were not undertaken. Specifically, this study attempted to answer the following research questions:

Can active or cooperative learning be successfully implemented using R to teach on the topic of sampling distributions in introductory statistics classes? If so, how can the program be used in active learning?

What are students general reaction to the program or its use in instruction? Since R is a harder program to master, do students react negatively to its use or do they accept its use?

What differences exist in cognitive measures on sampling distributions between R and other approaches (controlling for prior general statistics knowledge)?

1. Theoretical Framework

The question of how a student best learns statistics has been considered much in research (e.g., Garfield 1995; Gal and Garfield 1997; Lovett and Greenhouse 2000; Chance 2005). Garfield (1995) suggests that students learn best by constructing knowledge and becoming active participants in the learning process. Research in statistical instruction has mainly focused on instructional content or methods. In terms of instructional content, many statisticians, including Bradstreet (1996) and Cobb (1991), are convinced that an introductory statistics course should emphasize data analysis over mathematical technique and concepts over formulas. Hogg (1991) stressed that statistics should not be presented as a math course at all. Rather, the pedagogy should emphasize statistical reasoning and thinking rather than algebraic precision.

Smith (1998) indicated that: "One way to help students develop their statistical reasoning is to incorporate active-learning strategies that allow students to supplement what they have heard and read about statistics by actually doing statistics -- designing studies, collecting data, analyzing their results, preparing written reports, and giving oral presentations." (¶ 3.7). In support of an active learning approach, Bradstreet (1996) writes that, "Learning is situated in activity. Students who use the tools of their education actively rather than just acquire them build an increasingly rich implicit understanding of the world in which they use the tools and of the tools themselves." (pp. 73-74).

There are a variety of ways in which to incorporate active learning and projects into instruction. These might include some of the following: computer simulations (Garfield and delMas 1991); laboratory-based course (Bradstreet 1996); in-class activities (Dietz 1993; Gnanadesikan, Scheaffer, Watkins, and Witmer 1997); a single three-week project (Hunter 1977); or a course-long project (Fillebrown 1994; Mackisack 1994; Ledolter 1995; Chance 1997). At the core of most of these suggestions is the use of computers.

In regard to cooperative learning, many researchers have reported significant accomplishments from introducing cooperative learning experiences in introductory statistics classes (Shaughnessy 1977; Jones 1991; Dietz 1993; Keeler and Steinhorst 1995). Teams help to encourage cooperative learning, develop team-working skills, and usually build substantial friendship (Smith 1998). However, much of the research available tends to limit such activities to external learning situations (such as homework or studying), especially if computers are involved due to the rising cost of statistical software or tutorial programs. Furthermore, the use of computers in cooperative learning are often limited due to licensing issues of the software (e.g., students only being able to use the software in a lab) and the availability of computers.

Johnson and Dasgupta (2005) found that undergraduate students predominately like non-traditional instructional styles. Surprisingly, the use of computers in teaching statistics is still found to be a non-traditional instructional style due to the software and computer restrictions detailed previously. Exactly how a class should be structured and which techniques work best with each other is of major concern to most instructors. Jordan (2007) stressed that the implementation of such instructional styles is open to much interpretation. Lovett (2001) said that "a successful route to improving students' transfer of statistical reasoning skill may rely heavily on integrating instructional and cognitive theory, while maintaining a link to the realities of the classroom" (p. 348). Some research has considered the effect of combining different instructional techniques to create a "hybrid" approach in teaching statistics. Ward (2004) created a "hybrid" class consisting of online and face-to-face classes and found little difference in student performance. Keeler and Steinhorst (1995), and Steinhorst and Keeler (1995) created a "hybrid" class consisting of collaborative groups and mini-lectures. Their research focused on undergraduate, introductory-level statistics classes. Vaughn (2006) extended this idea for a graduate-level statistics class by merging three streams of instruction along with computer activities: lecture, active-learning in the classroom, and collaborative learning outside the classroom. Keeler and Steinhorst, Steinhorst and Keeler, and Vaughn showed an improvement in students' attitudes and grades when incorporating more active student involvement with lectures.

In teaching sampling distributions and Central Limit Theorem with technology, a variety of approaches have been considered. Ben-Zvi (1997) categorized computer approaches in one of the following ways: modified professional statistical packages, dedicated tutorials, and microworlds (which enable more interactive experimentation and simulation). Garfield, Chance, and Snell (2000) further delineate the uses of technology in college statistics classes by adding calculators, internet applets, programming languages, and spreadsheets (e.g., Excel) as useful instructional means. Doane (2004) incorporated an Excel-based approach in teaching distributions via simulation, although not exclusively sampling distributions. Most of the statistical tutorial software contain modules on these topics, and some actually focus exclusively on this topic (e.g., *Sampling Distribution* (delMas 1997)). Garfield et al. actually promote the *Tools for Teaching and Assessing Statistical Inference* project (Garfield, delMas, and Chance 2001) as an exemplary approach over others since it provides more detail and flexibility.

In regard to empirical research, Mills (2002) provides a review of the literature in using computer simulation methods to teach statistics and includes details on which studies contain empirical research. Although not exhaustive, this resource provides a nice starting point on the issue. Lipson (1997) compared two computer program (Minitab and Sampling Laboratory) via concept maps of students and showed that both approaches were equally effective. The pedagogical differences in each program suggest that the visual aspect in learning was more important than other matters such as simulated versus real data. delMas, Garfield, and Chance (1999) used the *Sampling Distribution* program to test for significant changes in student understanding. Their results showed a significant increase in post-test scores, yet the authors were still disappointed in fundamental understandings of the Central Limit Theorem by the students of the study. They concluded that software alone is insufficient to impart knowledge of this concept. Lunsford, Rowell, and Goodson-Espy (2006) used a mixed-methods research design to look at students' understanding of sampling distributions. Their finding indicated that instructor classroom demonstration of sampling distributions and the Central Limit Theorem was not enough for effective impartation of knowledge with students. Rather, students needed hands-on active-learning opportunities to fully grasp these ideas.

The numerous research listed above indicates that the combination of both visual instruction and active learning stimulates comprehension. The exact nature of this active learning can take on various forms, as can the technology or software used to aid in this activity. While most studies listed above detail the use of JAVA applets or traditional software packages, few resources exist that give strategies for incorporating R into classes. This article will detail one approach to use R to actively teach sampling distributions.

The uniqueness of the approach taken in this article is that use of R highlights many of the strengths of other technological approaches. Using Ben-Zvi's (1997) categorizations, the R program can be viewed as a hybrid approach. That is, it is a professional statistical package that allows for interactive experimentation and simulation, and thus could be used as a dedicated tutorial. No modification of the program is necessary to use it in this capacity. Using Garfield, Chance, and Snell's (2000) recommendations of software that provides more detail and flexibility, the use of R allows for both easily. Instructors are able to modify the approach of this article to give less/more detail for students. For example, some instructors may wish to incorporate discussion of confidence intervals after presenting instruction on sampling distributions and the Central Limit Theorem, and this addition could be easily incorporated with R. An added benefit for students is that any instructional use of R will help them to become familiar with an actual professional statistical analysis program that they could use in their own research or profession.

In addition, the use of R allows an instructor to customize the amount of interaction the student has in the activity. For example, it would be easy to customize the presentation in this article to allow for a more constructivist opportunity in learning. Team work could easily be allowed in this approach. The instructor can quiz students at different stages to enforce the active component of the exercise. Each stage of the activity could be explored in detail to expand the intermediate results. The key in all of these variations is that use of R allows this flexibility. The disadvantage of applets is that often you are left with the author's vantage point for how the concept should be taught and there is little flexibility in revising it according to a particular teaching paradigm. Also, many applets have limitations of dealing with one simulation condition at a time. For example, it is usually difficult to compare the effects of changes in sample size on sampling distributions with most applets unless the results are printed off or memorized by the students.

2. Teaching Goals

The main goal in this approach is to allow students an opportunity to discover various concepts related to the Central Limit Theorem, in addition to understanding the nature of sampling distributions. As such, the presented approach allows the student or user direct manipulation of the simulated data. This activity can be modified to cover less/more interaction as deemed important by the instructor.

While many of the demonstrations using R could be streamlined (e.g., written as a function so that students input a few numbers to get the end result), the goal was to design activities which enabled more hands-on approach to learning for the student. Thus, this article details a more laborious approach that hopefully will enable students to understand the steps involved and a better understanding of the emerging concept. The approach in the article assumes a certain understanding of students in regard to R (e.g., in my course, my students will have used R extensively up to this point). Thus, if an instructor does not use R regularly, there will be need for more preparatory work with students before the approach in this article can be used.[1] At this point of instruction, students should have already been exposed to ideas of distributions, central tendency, variability, and so on. Students should also be exposed to the idea of sampling and simulation.[2]

Additionally, the approach taken in this article is but one of many ways to do this in R. Only one particular example is demonstrated. The purpose of the article is not to suggest that the presented coding is the only way to present this material, nor the "best" way. Nor is the illustration meant to be exhaustive of the ideas surrounding sampling distributions. Rather, the purpose is to provide a beginning

[1]Instructors can certainly use R for active learning without using the program for data analysis. Students may find the program difficult to use if this is their first exposure to it. Fortunately, as mentioned in the article, an instructor can customize the code to enable more ease for the student (at a cost of less interactivity though).

[2]Both of these were true for the students in this study. For example, students in this study did hands-on experiments with sampling and simulation in the unit covering probability (which was before the unit on sampling distributions).

point for instructors who want to experiment with the use of R in instruction (i.e., whether R be effectively used in this manner). Thus the code provided in this article is but one of many ways that an instructor might pursue the use of R in teaching statistics and are meant to be modified and changed to match a particular instructor. At times, some possible modifications will be suggested in footnotes (yet these modifications are not exhaustive). For instance, this activity was constructed as an introduction to the topic. As such, simple population distributions are used instead of real data in an attempt to not overwhelm students. A follow-up activity repeats this activity with real data. This follow-up activity is not detailed in this study. An instructor is encouraged to expand this example to include more interaction with students as they actively construct sampling distributions.

3. Logistics

3.1 The problem

One of the advantages in using R to teach is that one can easily carry on several examples at one time. The author will illustrate this for a traditional experiment of sampling distributions from a uniform distribution.[3] Let's assume that our population under consideration has the following distribution (Figure 1).

The variable *X* is discrete, and uniformly distributed. The mean of this population is 3, and the standard deviation is $\sqrt{2}$.[4] We will simultaneously see the effect of taking different sample sizes in R from this population. Let's consider the case of $n = 2, 10, 30, 50$, and 100.

Figure 1. Uniform Distribution Used for Classroom Simulation

[3] The uniform distribution is presented as a model of a fair 5-sided die. A traditional 6-sided die is sometimes used as well.

[4] I take my students through more of the details of these summaries. Since the article deals more with the use of R in teaching, I forgo some of the detailed instruction I use and rather focus exclusively on the R content from here on.

3.2 Setting Up the Simulation in R

There are many ways to simulate a uniform distribution in R, but we will use the following command: as.integer(runif(*n*,*min*,*max*)). The *sample* function could also be used,[5] yet this is not the approach taken in this article. The *runif* command will generate *n* numbers greater than the *min* and smaller than the *max* values. The *as.integer* command will round this number down to a whole number, thus allowing us to mimic discrete distributions. For the example above, the author will be utilizing the command: as.integer(runif(*n*,1,6)). Note that we must increase the maximum number by one since the random numbers will be between the min and max, and all numbers will be rounded down. To simulate the sample sizes mentioned above, we will assign the average of each sample to a vector of results named *mean*. So, *mean 2* will represent the mean of samples of size 2 taken from this population, *mean10* will represent the mean of samples of size 10 taken from this population, and so on.

The first step to carrying this out in R is to define the dimensions for each of these vectors. The author will essentially carry out the sampling 10,000 times (although you can vary this). The screen in R will look like Figure 2.

3.3 Sampling Distributions in R

Now that the author have defined the vectors, we will carry out the simulation with the following commands[6] (Figure 3).

[5] For example, a "population" of values could be simulated using the commands in this article. The *sample* function could then be used to generate random samples from this "population." Certainly this is a viable alternative to what is shown in this article. This approach would be useful if one wanted students to be able to graph a histogram of population values and compare them to the sample values (however, one could also graph densities in the presented approach and achieve the same results). The approach in this article is useful to demonstrate tendencies in sampling distributions derived from general ideas of populations. Alternatively, a large population can be randomly generated, and then the *sample* function used to draw random samples (without replacement) from this (say 10% of the population).

[6] When I teach this, I take students through a "build up" period where we talk about each piece of the code. For example, we will first experiment with the *runif* function to see what it does. We will use it to generate a few values, then we will increase this and consider histograms and summary statistics from its use. Once they are comfortable with this idea, we will add the *as.integer* command and repeat the process. I continue this until students are comfortable with all coding before the simulation is done. The advantages in using R is that instructors have individualized choice for how much of this in-depth coverage they want for their students. Some instructors may wish their students to simply "copy/paste" the commands and focus on the results. This, along with coding the commands into an R package, is certainly another option available when using R.

```
> means2 = numeric(10000)
> means10 = numeric(10000)
> means30 = numeric(10000)
> means50 = numeric(10000)
> means100 = numeric(10000)
```

Figure 2. Setting Up Vectors for Simulation in R

```
> for(i in 1:10000){means2[i] = mean(as.integer(runif(2,1,6)))}
> for(i in 1:10000){means10[i] = mean(as.integer(runif(10,1,6)))}
> for(i in 1:10000){means30[i] = mean(as.integer(runif(30,1,6)))}
> for(i in 1:10000){means50[i] = mean(as.integer(runif(50,1,6)))}
> for(i in 1:10000){means100[i] = mean(as.integer(runif(100,1,6)))}
```

Figure 3. Simulation for Five Sample Size Conditions

For each value of *i* in our simulation of 10,000 samples, the *i*th value in the vector is replaced with the mean of that particular sample. Each simulation condition takes only one to two seconds to run. Please note that the simulation condition shown above is contained within {} brackets and not parentheses.[7] For instructional purposes, it might be beneficial to build the commands shown in Figure 3 in stages. That is, one might have students try using the *runif* function several times in a row to see what this command does (hitting the up arrow on the keyboard will bring up the last typed command in R which makes repeating this "mini-experiment" much easier). Next, students can try out the additional *as.integer*[8] command to see what this command adds to the previous one, and so on. To plot the histogram of results from any simulation condition, simply use the function *hist(vector name)*. For example, to obtain the histogram for *means10*, one would type *hist(means10)* and get the following histogram from R:

Obviously, this is based on this particular sample, and each student will have a slightly different histogram. One can incorporate a *seed* function to fix the random numbers to be alike for each student if so desired.

3.4 Comparing Conditions in R

The author often quiz students at this point to see what they expect to happen in the results before one graph them. The real benefit in running several simulation conditions at once is in comparing the histograms for the multiple simulations we have run. It would be nice to see them in progression so one can visualize what happens as n increases. One way to do this in R is to use the par command which will enable one to graph multiple graphs in a matrix of columns and rows. Since the study have 5 conditions in this simulation, let's construct a matrix of 2 columns, and 3 rows to see the histograms:

A blank window should open up in R. The graphs will go in this window. It would be beneficial to put all of the histograms on the same scale as the population the study sampled from, so for each histogram[9] we will deliberately ask for a particular minimum and maximum value:

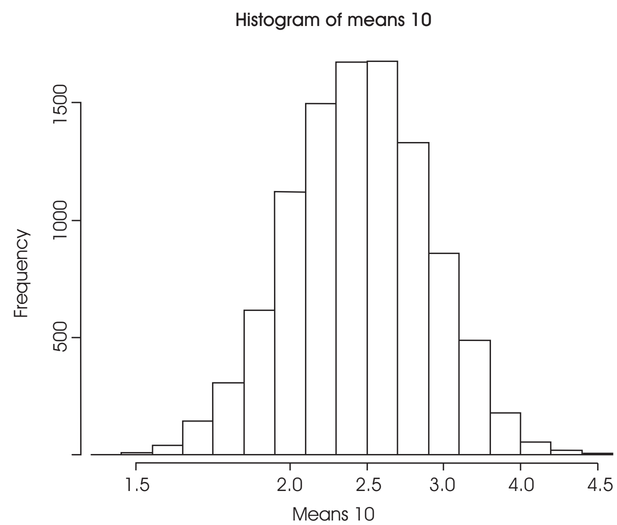

Figure 4. Histogram for Simulation Involving Sample Size of 10

```
> par(mfrow = c(3,2))
```

Figure 5. Setting Up a Matrix for Multiple Graphs in R

```
> hist(means2, xlim = c(1,5), ylim = c(0,3000))
> hist(means10, xlim = c(1,5), ylim = c(0,3000))
> hist(means30, xlim = c(1,5), ylim = c(0,3000))
> hist(means50, xlim = c(1,5), ylim = c(0,3000))
> hist(means100, xlim = c(1,5), ylim = c(0,3000))
```

Figure 6. Plotting Histograms with Equal Axes

[7] An alternative to the *for* statement is the use of *repeattrials*. This has the advantage of avoiding *for* loops in R and might be preferred by some users. So instead of the *for* statements in Figure 3, one could use a statement like this instead (illustrated for the case of sample size of 5): *means5 = repeattrials (mean(as.integer(runif(5, 1, 6))), 10000).* However, this function is available in a package which must be installed and loaded into R before use (see http://www.stat.psu.edu/~dhunter/016/repeat.r for more information). I typically minimize having introductory statistics students load in created functions, so this approach is not emphasized in this paper.

[8] An alternative to the *as.interger* command for rounding down is the *floor* command. For example, the first command could be rewritten as: *for (I in 1:10000) {means2[i] = mean(floor(runif(2,1,6)))}*.

[9] An alternative to this approach is graphing densities instead of histograms. To illustrate this for the $n = 50$ and 100 conditions, one could type: *plot(density(means100))* followed by *lines(density(means50), col='red')*. Notice this approach overlays the graphs and additionally changes the color of the second graph to make it easier to compare. One could also plot the densities in separate graphs using the *par* command as shown in the article. The use of the *density* option can be beneficial as it would be possible to graph the density of the population and compare this to the sampling distributions. If the *density* option is used, it is recommended to start with the largest sampling condition first and work backwards so that the axes scale will be correct.

This would result in the following graph (Figure 7):

Since the study have stored each simulation condition in a unique vector, we can easily see descriptive statistics for these distributions (Figure 8). One could ask more means, summaries, and standard deviations for each condition separately, but a shortcut is presented instead in Figure 8. All vectors are combined into a common dataset (called *data* in this example). Now, we can simply request a summary of the data (which will give us the mean), along with the standard deviation (using the *sd* function). Thus, this approach shows calling for all summary statistics at once. However, some students might find it easier to just call up descriptive statistics for each condition using the *mean* and *sd* commands in R (e.g., *Sd(means2)*).

If one would like to compare these standard deviations of the means (i.e., the standard error) to the theoretical derivation from the Central Limit Theorem (n), they can use the command shown in Figure 9. Recall that the standard deviation for our population under question was

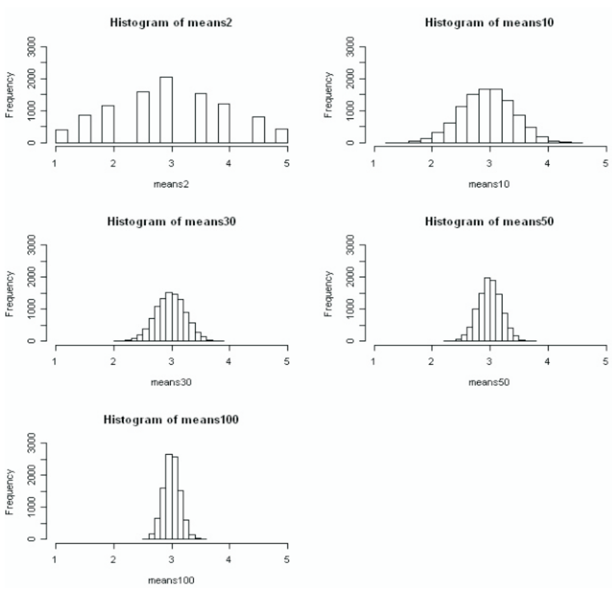

Figure 7. Matrix of Histograms for Different Sample Sizes of Simulation

```
> data = cbind(means2, means10, means30, means50, means100)
> summary(data)
     means2          means10          means30          means50          means100
 Min.   :1.000   Min.   :1.300   Min.   :2.067   Min.   :2.260   Min.   :2.530
 1st Qu.:2.500   1st Qu.:2.700   1st Qu.:2.833   1st Qu.:2.860   1st Qu.:2.910
 Median :3.000   Median :3.000   Median :3.000   Median :3.000   Median :3.000
 Mean   :3.004   Mean   :2.998   Mean   :2.996   Mean   :3.002   Mean   :3.002
 3rd Qu.:3.500   3rd Qu.:3.300   3rd Qu.:3.167   3rd Qu.:3.140   3rd Qu.:3.090
 Max.   :5.000   Max.   :4.600   Max.   :3.833   Max.   :3.760   Max.   :3.580
> sd(data)
   means2    means10    means30    means50   means100
1.0079981 0.4509983 0.2540334 0.1993863 0.1417976
```

Figure 8. Sample Statistics for the Means from Different Sample Size Conditions

```
> s.size = c(2,10,30,50,100)
> sqrt(2)/sqrt(s.size)
[1] 1.0000000 0.4472136 0.2581989 0.2000000 0.1414214
```

Figure 9. Theoretical Values of Standard Error to Compare to Simulation Values

2 . Figure 9 shows the creation of a vector called *s.size* which contains the values of the sample size for each simulation condition. Constructing it this way allows for a simultaneous calculation of the theoretical standard error.

4. Actual Use

4.1 Student Reaction 3

Students are presented this routine through a classroom demonstration and a series of free online instructional videos to reinforce the procedure in R (author link provided later due to blind review process)[10] After watching the videos, students then work through their own examples (via a portfolio method) or a provided worksheet detailing various assigned activities at their convenience. This method has worked in an exemplary manner due to R being a free program package that students can download. Students have indicated that they understand the material much better after having worked through different distributions.

In varying semesters, the author have used Java applets, free dedicated sampling distribution programs (e.g., Sampling SIM (delMas, 2002)), and most recently R. he found, surprisingly, that students struggled more with the Java applets and free dedicated programs perhaps due to the fact that the applet's content did not always match up to the way he taught the subject. Students also struggled with the massive amounts of information that were on the screen at any given time. By using R, he was able to customize the computer application around the way he taught the topic.

In a recent semester, the author used an introductory statistics class that used R to compare to other classes that did not get this treatment. He interviewed a random sample of students about R from that class. One student

[10] The videos cover all major topics in Intro Stat. Students are not required to watch the videos, and could watch the videos out of sequence if they choose. At various points in the semester, students are encouraged to watch a particular video covering the topic being taught.

remarked, "I wasn't sure about this program because it looked hard. But when I actually sat down to use it, I found out that it was really easy. And I like the program!" Another student commented, "R is so cool!" One student commented about how she was leery of the program at first: "I was used to other programs so I wasn't too happy about using R. But when I really gave it a chance, I saw what you meant about how it is much easier to use than it looks. Of course, I know that a lot of that was your instructions and online videos. Those helped a lot!" In this particular sample, the author found no student with a negative critique of R. Most students were extremely enthusiastic about it. Of the students who were more neutral about the program, it seemed that their neutrality dealt more with their view of their future use of statistics and not the R program itself.

Another interesting result of using R is that email questions regarding exercises actually decreased. For example, in semesters that the author used online Java applets or Sampling SIM, he would constantly have frequent emails for help. When he used R, this frequency was dramatically decreased. However, this effect is obviously confounded with the online videos he have made available to students which were not available for the applets. Thus, it is hard to discern whether the availability of videos or the R program itself decreased this frequency of questions and problems.

Collectively, it appeared that students reacted to the program favorably and accepted its use in their instruction. The negative reactions expected from the use of R were not present in this class.

4.2 Empirical Comparison

To test the effectiveness of a free dedicated sampling distribution program (Sampling SIM) to an R-code approach, different introductory statistics classes were compared. The classes consisted of mostly undergraduate students at a major university in the south. The course is offered through the Educational Psychology Department in the College of Education and is taken by a variety of students from education, social work, nursing, music, communications, and so on. There is no prerequisite for the class.

Assessment of student knowledge on sampling distributions was considered as the cognitive measure of the effectiveness of each instructional method. The cognitive measure was taken from a master list of questions which was developed by the researcher over a 10-year period. The master list contains 100 multiple choice questions, each with three distracters and one correct answer, covering topics in sampling distributions and the Central Limit Theorem. A secondary set of 212 items (covering more elementary concepts such as graphs and descriptive statistics) was used in this study as a covariate. The same cognitive measure (items dealing with sampling distrubitons) was used in all three classes since students were not allowed to keep the measure. Psychometric issues, such as validity and reliability, have been considered on these items. Face validity was conducted with other statistics instructors deemed as content experts, and the validity was found exemplary. Construct validity has been carried out in various analyses over the course of development, as well as reliability assessment and item analysis. Over a period of years, items have been removed, adjusted, or improved upon. This final bundle of items has shown to have sound psychometric properties of consistency (coefficient alpha's at least 0.80) and factor loadings consistent with the construct that the item is measuring. Each assessment per class consisted of the same 33 randomly selected questions from this bank of questions. The smallest coefficient alpha for the pre- and post-measure in this study was 0.82 and 0.84, respectively.

Each class was taught by the same instructor over three semesters. To help limit possible confounding due to semester effects and non-random sampling, the same

		Prior Measure Cognitive		Final Measure Cognitive	
	N	Mean	SD	Mean	SD
Control	66	85.62	10.52	80.94	15.00
Sampling SIM	76	88.64	8.81	83.29	9.48
R Code	102	83.51	9.10	90.10	11.24
Total	244	85.84	9.75	85.50	12.51

Table 1. Descriptive statistics for prior and final cognitive measures

pre-test measure was given in each class (approximately 2 weeks before the activity). Table 1 summarizes the cognitive measures for each class (on a 100 point scale).

A between-subjects analysis of covariance was performed on the final cognitive measure. The independent variable consisted of type of instruction (Control (no computer component), *Sampling SIM*, R code). Each class worked through a similar exercise constructing sampling distributions. The control group constructed sampling distributions by hand. The *Sampling SIM* applet class constructed sampling distributions via an applet. And the R Code group used the procedure illustrated in this article. The covariate was a prior cognitive assessment of knowledge over basic statistical knowledge (graphs, descriptive statistics, and so on).

An initial check of assumptions revealed no problems with outliers and normality. A boxplot for each group is given in Figure 10. Since the cognitive outcomes were measured in such a way as to be considered a proportion, an analysis was also considered in which the cognitive measures were transformed using an arcsin transformation as suggested by Neter, Kutner, Nachtsheim, and Wasserman (1996). Using these transformed variables, the assumptions of normality of sampling distributions, linearity, homogeneity of variance,

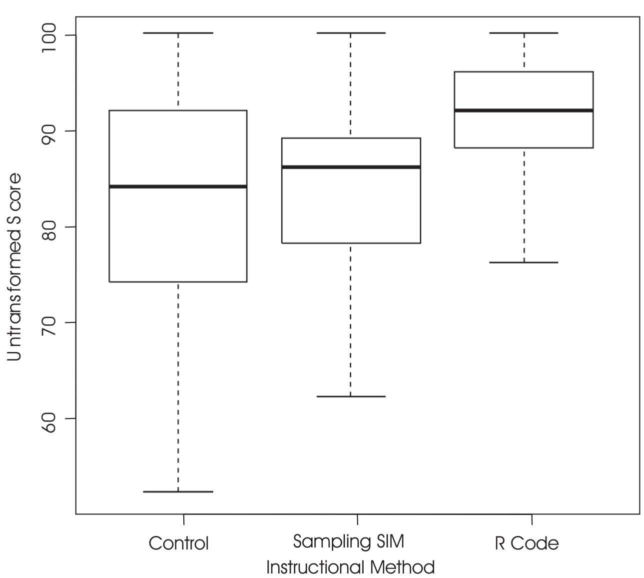

Figure 10. Boxplots of final cognitive measure by instructional methods

homogeneity of regression, and reliability of the covariate were found to be satisfactory. However, the results were consistent with that of the untransformed variables. The ANCOVA results are presented in terms of the transformed variables; yet for ease in interpretation, group means are presented in terms of the original scale.

After adjusting for prior cognitive abilities, final cognitive assessment varied significantly with type of instruction ($F(2, 236) = 11.387$, $p < 0.001$). The ANCOVA results are summarized in Table 2. The strength of the relationship between adjusted cognitive final measure and type of instruction was slightly moderate with partial = 0.088. The adjusted non-transformed marginal means, as shown in Table 3, indicate that the highest final cognitive measure was for the class that used the R code. The lowest final cognitive measure was for the class with the non-computer activity. A post-hoc test revealed significant differences between the "R Code" based approach and the control group, as well as between the "R Code" based approach and the "*Sampling SIM* applet" approach. There were no significant differences between the "*Sampling SIM* applet" approach and the control group based approaches. Based on these results, students who use R over other methods have significant gains in cognitive understanding in sampling distributions, on average.

The empirical research in this study is limited by such issues as random sampling and assignment, as well as breadth in comparisons to other procedures. Only one particular

Source of Variance	Adjusted SS	df	MS	F
Type of instruction	1.616	2	0.808	11.387*
Covariate (adjusted for all effects)				
Prior cognitive measure	2.186	1	2.186	30.807*
Error	16.746	236	0.071	

Table 2. Analysis of Covariance of Final Cognitive Measure

Type of Instruction	Adjusted Mean	Unadjusted Mean
Control	82.65	80.94
Sampling SIM	85.59	83.29
R Code	89.51	90.10

Table 3. Adjusted and Unadjusted Mean Cognitive Measure for Three Designs of Instruction (untransformed)

free applet was used to compare to R. Other cognitive measures should be considered (as well as non-cognitive measures). More research should be done in this area to replicate the results of this study to see if in fact the use of R does provide an actual advantage for classroom instruction. The evidence does suggest that R can effectively be used in active learning situations. The results do not suggest that R is the "best" approach to use. Rather, given the limitations of the study, a better conclusion would simply be that the use of R is effective. With just one study, the conservative conclusion of the empirical data would be to say that the use of any computer component to active learning is indeed beneficial and that in particular the use of R appears as good (possibly better for the instructor in this study) than free applets. Whether the results would replicate to other instructors is a matter of future study.

4.3 Benefits

The obvious benefit of this approach to the teacher is in cost and time. While various instructional components exist in the form of Java applets or commercial products, this approach has no direct cost associated with it and can be customized at will by the instructor. For example, if an instructor only wants their students to grapple with the idea of a sampling distribution without considering the resulting shape or properties, this can easily be done with this approach. This might be useful if the instructor would like their students to arrive at the Central Limit Theorem through a more constructivist approach (versus a tutorial which tells the student the results). Although this researcher utilizes the code via at-home assignments, other instructors could easily use this to only demonstrate the concept in class or to incorporate active-learning components via a computer lab. Instructors are free to use the above link to train their students in constructing sampling distributions in R (and learn about R in general), thus saving even more time.

4.4 Implementation

R has a shortcut which makes implementation of this code easier. Once a user enters a command, if the next command is similar the user can simply hit the "up" arrow on their keyboard and revise the previous command. This shortcut can be a tremendous time-saver for instructors who wish to implement this approach in a lab setting. For those instructors who have severe limitations, a file with all the commands can be provided to the students as shown in this file (author link provided after blinded review). Students can simply copy and paste the code into R to run the commands. Students can modify the code as needed for multiple active-learning lessons (e.g., changing the distribution from which samples are taken).

Any instructor who is used to teaching with technology will find that using R requires about the same prep time. The pros of using R is that once the code is written, it can be used over again, or easily modified to fit new teaching emphasis. The newness of R may be a con for some instructors who are not used to the program. Thus, if an instructor is not used to R and decides to incorporate it in teaching, higher prep time should be expected. Using R for data analysis is advised first before attempting to use the program interactively in class.

5. Extensions

One of the main extensions of this activity in R would be to vary the population from which samples are drawn. Often, one would like to show both discrete and continuous populations, in addition to varying degree of skewness versus symmetry. Some of the more common

Distribution	R Command	Useful for
Uniform – Discrete	as.integer(runif(*sample size*, *min*, *max+1*))	Discrete population densities which are flat in shape within range of *min* and *max*
Uniform – Continuous	runif(*sample size*, *min*, *max*)	Continuous population densities which are flat in shape within range of *min* and *max*
Binomial	rbinom(*sample size*, *number of trials*, *probability of success on each trial*)	Discrete population densities useful in situations where each trial has only dichotomous outcomes
Poisson	Rpois(*sample size*, *lambda*)(	Discrete population densities for phenomena that may happen 0, 1, 2, 3, ... times during a given period of time
Normal	rnorm(*sample size*, *mean*, *standard deviation*)	Continuous population densities which are symmetrical, bell-shaped
Beta	rbeta(*sample size*, *shape1*, *shape2*)	Continuous population densities which can model skewed distributions

Table 4. Examples of Different Distributions Available in R

distributions in R are summarized in Table 4. These commands could be used in place of the discrete uniform distribution demonstrated in Figure 3.

Other adjustments could also be made, such as changing the sample size per iteration, and the number of iterations. These adjustments should be easy to make and are not detailed. Also, real data could be utilized in addition to the simulated data, or a large population could be simulated with the above commands and the *sample* function used to draw a random sample (without replacement) from these populations.

Different sampling distributions (e.g., proportion) can be easily incorporated in the presented code. The intention of this article is to provide a basis for performing such an experiment with students. Also, more automation can be done by the creation of functions within R. For example, one could write a function that would perform the entire simulation and output based on the code provided. However, as previously mentioned, much research indicates that students might understand the concept more by considering the step-by-step process that is involved in creating sampling distributions. Also, the provided code could be re-written to allow for one matrix containing all simulation conditions. While this would provide parsimony in the code, this researcher has found that "Introductory Statistics" students struggle more with this approach to coding than in treating each condition as a separate vector. Certainly for more advanced statistics students who are familiar with matrix algebra, this recoding might prove beneficial.

Conclusions

This study considered whether the program R could be used to teach sampling distributions from an active learning perspective, as well as how students reacted to this both affectively and cognitively. One approach (utilizing common population distributions) was demonstrated in this article. This approach is one of many that R could be used for. As mentioned in the article, one could customize this code to emphasize random samples from real data, random samples from large populations (which are randomly generated), or random samples from unknown populations. The author actually uses combinations of all approaches mentioned above in several activities. The only approach presented in the article is the more traditional way that sampling distributions are taught.

Due to the limitations in the design of this study, the generalization of results should be done with caution yet are worthy of consideration, nonetheless. Since the same instructor was used in all samples, there could be some confounding of the results. Yet, the results showed for this instructor, students were not as fearful of the R program as was initially feared by the instructor. Futhermore, the R program appeared to be as affective in assisting sampling distribution instruction as Java applets (and better than not using any technology). This is encouraging since the use of R showcases a different paradigm than JAVA applets (i.e., a more dynamic, customizable approach that can be tailored around an individual's mode of instruction).

This study provides one example of how R could be used to actively teach sampling distributions. Further research can help amplify the various uses for R and even compare various approaches to teaching the same concept using R.

References

[1]. **Ben-Zvi, D. (1997),** "Discussion: Software for Teaching Statistics," in *Research on the Role of Technology in Teaching and Learning Statistics*, eds. J. Garfield and G. Burrill, Voorburg, The Netherlands: International Statistical Institute, pp. 123-136.

[2]. **Bradstreet, T. E. (1996),** "Teaching Introductory Statistics Courses So That Nonstatisticians Experience Statistical Reasoning," *The American Statistician*, 50, 69-78.

[3]. **Chance, B. L. (1997),** "Experiences with Authentic Assessment Techniques in an Introductory Statistics Course," *Journal of Statistics Education*, [Online], 5(3). (http://www.amstat.org/publications/jse/v5n3/chance.html)

[4]. **- - (2005),** "Integrating Pedagogies to Teach Statistics," in *Innovations in Teaching Statistics*, ed. J. Garfield, MAA Notes volume #65.

[5]. **Cobb, G. W. (1991),** "Teaching Statistics: More Data,

Less Lecturing," *Amstat News*, December, No. 182, 1 and 4.

[6]. **Cohen, S., and Chechile, R. A. (1997),** "Overview of ConStats and the ConStats Assessment," in *Research on the Role of Technology in Teaching and Learning Statistics*, eds. J. Garfield and G. Burrill, Voorburg, The Netherlands: International Statistical Institute, pp. 99-108.

[7]. **Crawley, M. J. (2005),** *Statistics: An Introduction Using R*, West Sussex, England: John Wiley & Sons Ltd.

[8]. **delMas, R. C. (1997),** "A Framework for the Development of Software for Teaching Statistical Concepts," in *Research on the Role of Technology in Teaching and Learning Statistics*, eds. J. Garfield and G. Burrill, Voorburg, The Netherlands: International Statistical Institute, pp. 75-90.

[9]. **delMas, R. C. (2002),** *Sampling SIM*, version 5.4 [On line]. www.gen.umn.edu/research/stat_tools/

[10]. **delMas, R. C., Garfield, J., and Chance, B. L. (1999),** "A Model of Classroom Research in Action: Developing Simulation Activities to Improve Students' Statistical Reasoning," *Journal of Statistics Education* [Online], 7(3). (www.amstat.org/publications/jse/secure/v7n3/delmas.cfm)

[11]. **Dietz, E. J. (1993),** "A Cooperative Learning Activity on Methods of Selecting a Sample," *The American Statistician*, 47, 104-108.

[12]. **Doane, D. P. (2004),** "Using Simulation to Teach Distributions," *Journal of Statistics Education*, [Online], 12(1).(http://www.amstat.org/publications/jse/v12n1/doane.html)

[13]. **Fillebrown, S. (1994),** "Using Projects in an Elementary Statistics Course for Non-science Majors," *Journal of Statistics Education*, [Online], 2(2). (http:// www.amstat.org/publications/jse/v2n2/fillebrown.html)

[14]. **Gal, I., and Garfield, J. B. (Eds.) (1997),** *The Assessment Challenge in Statistics Education*, Amsterdam: IOS Press / International Statistical Institute.

[15]. **Garfield, J. (1995),** "How Students Learn Statistics," *International Statistical Review*, 63(1), 35-48.

[16]. **Garfield, J., Chance, B., and Snell, J.L. (2000),** "Technology in College Statistics Courses," in *The Teaching and Learning of Mathematics at University Level: An ICMI Study*, eds. D. Holton, et al, The Netherlands: Kluwer Academic Publishers.

[17]. **Garfield, J. and delMas, R. C. (1991),** "Students' Conceptions of Probability," in *Proceedings of the Third International Conference on Teaching Statistics, Volume 1*, ed. D. Vere-Jones, Voorburg, The Netherlands: International Statistical Institute, pp. 340-349.

[18]. **Garfield, J., delMas, R. C., and Chance, B. (2001),** "Tools for Teaching and Assessing Statistical Inference," paper presented at the Joint Mathematics Meetings, New Orleans, LA.

[19]. **Gnanadesikan, M., Scheaffer, R. L., Watkins, A. E., and Witmer, J. A. (1997),** "An Activity-based Statistics Course," *Journal of Statistics Education*, [Online], 5(2). (http://www.amstat.org/publications/jse/v5n2/gnanadesikan.html)

[20]. **Hogg, R. V. (1991),** "Statistical Education: Improvements are Badly Needed," *The American Statistician*, 45, 342-343.

[21]. **Hunter, W. G. (1977),** "Some Ideas About Teaching Design of Experiments, with 2 ^ 5 Examples of Experiments Conducted by Students," *The American Statistician*, 31, 12-17.

[22]. **Johnson, H. D., and Dasgupta, N. (2005),** "Traditional Versus Non-traditional Teaching: Perspectives of Students in Introductory Statistics Classes," *Journal of Statistics Education*, [Online], 13(2). (http://www. Amstat. Org/publications/jse/v13n2/johnson.html)

[23]. **Jones, L. (1991),** *Using cooperative learning to teach statistics. Research Report Number 91-2*, University of North Carolina: The L. L. Thurstone Psychometric Laboratory.

[24]. **Jordan, J. (2007),** "The Application of Statistics Education Research in My Classroom," *Journal of Statistics Education*, [Online], 15(2). (http://www.amstat.org/publications/jse/v15n2/jordan.html)

[25]. **Keeler, C. M., and Steinhorst, R. K. (1995),** "Using Small Groups to Promote Active Learning in the Introductory Statistics Course: A Report From the Field," *Journal of Statistics Education*, [Online], 3(2). (http:// www.amstat.org/publications/jse/v3n2/keeler.html)

[26]. **Ledolter, J. (1995),** "Projects in Introductory Statistics Courses," *The American Statistician*, 49, 364-367.

[27]. **Lipson, K. (1997),** "What Do Students Gain From Simulation Exercises? An Evaluation of Activities Designed to Develop an Understanding of the Sampling Distribution of a Proportion," in *Research on the Role of Technology in Teaching and Learning Statistics*, eds. J. Garfield and G. Burrill, Voorburg, The Netherlands: International Statistical Institute, pp. 137-150.

[28]. **Lovett, M. C. (2001),** "A Collaborative Convergence on Studying Reasoning Processes: A Case Study in Statistics," in *Cognition and Instruction: Twenty-five Years of Progress,* eds. S. M. Carver and D. Klahr, Mahwah, NJ: Erlbaum, pp. 347-384.

[29]. **Lovett, M. C., and Greenhouse, J. B. (2000)**, "Applying Cognitive Theory to Statistics Instruction," *The American Statistician*, 54(3), 196-206.

[30]. **Lunsford, M. L., Rowell, G. H., and Goodson-Espy, T. (2006),** "Classroom Research: Assessment of Student Understanding of Sampling Distributions of Means and the Central Limit Theorem in Post-calculus Probability and Statistics Classes," *Journal of Statistics Education*, [Online], 14(3). (http://www.amstat.org/publications/ jse/v 14n3/lunsford.html)

[31]. **Mackisack, M. (1994),** "What is the Use of Experiments Conducted by Statistics Students?" *Journal of Statistics Education*, [Online], 2(1). (http://www.amstat.org/publications/jse/v2n1/mackisack.html)

[32]. **Mills, J. D. (2002),** "Using Computer Simulation Methods to Teach Statistics: A Review of the Literature," *Journal of Statistics Education*, [Online], 10(1). (http://www.amstat.org/publications/jse/v10n1/mills.html

[33]. **Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996).** *Applied linear statistical models* (4th ed.). Boston, MA: McGraw-Hill.

[34]. **Shaughnessy, J. M. (1977),** "Misconceptions of Probability: An Experiment with a Small-group Activity-based Model Building Approach to Introductory Probability at the College Level," *Educational Studies in Mathematics*, 8, 285-315.

[35]. **Smith, G. (1998),** "Learning Statistics by Doing Statistics," *Journal of Statistics Education*, [Online], 6(3). (http://www.amstat.org/publications/jse/v6n3/smith.html

[36]. **Steinhorst, R., and Keeler, C. (1995),** "Developing Material for Introductory Statistics Courses from a Conceptual, Active Learning Viewpoint," *Journal of Statistics Education*, [Online], 3(3). (http://www.amstat.org/publications/jse/v3n3/steinhorst.html)

[37]. **Vaughn, B. K. (2006),** "Learning Educational Statistics by Doing Educational Statistics," paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 2006.

[38]. **Velleman, P. (1996),** *ActivStats*, Ithaca, NY: Data Description.

[39]. **Verzani, J. (2005),** *Using R for Introductory Statistics*, Boca Rotan, FL: CRC Press..

[40]. **Ward, B. (2004),** "The Best of Both Worlds: A Hybrid Statistics Class," *Journal of Statistics Education*, [Online], 12(3). (http://www.amstat.org/publications/jse/v12n3/ward.html)

ABOUT THE AUTHOR

Brandon Vaughn's current research interests include multi-level differential item functioning (DIF), Bayesian estimation procedures, creative uses of non-parametric classification procedures, and effective strategies in the teaching of statistics. He has developed several technological tools for teaching statistics, including free R tutorial videos and applets for conceptual understanding.